


# Open Access Statistics : an examination how to generate interoperable usage information from distributed open access services



Université Lille 3: International Symposium on  
„Academic Online Ressources : Assessement and Usage“  
26.11.2009

Initiated by:



**Ulrich Herb**

Saarland University and State Library, Germany  
u.herb@sulb.uni-Saarland.de

Funded by:

Deutsche  
Forschungsgemeinschaft

**DFG**

# overview

---

- impact measures: relevance
- impact measures: some categories
- usage based impact measures: standardization?
- DFG-Project: Open Access Statistics
  - motivation, associated projects, technical issues, some results
  - outlook

# impact measures: relevance

---

- individual level: *publish or perish*
  - a scientist that does not publish hardly has any reputation or impact
  - without any impact, he won't make his carrier
  
- organizational level: evaluation
  - evaluation results determine prospective resources of institutes *and* the future main research
  - criteria: number of doctoral candidates, amount of third party funds, publications

# from publications to impact

---

- scientific reputation (or scientific capital) is derived from publication impact
- impact is calculated mostly by citation measures
  - journal impact factor (jif)
  - hirsch-index (h-index)

*especially within the STM-domain*

# citation impact: calculation

---

## □ jif: calculation

in year  $X$ , the impact factor of a journal  $Y$  is the average number of citations to articles that were published in  $Y$  during the two years preceding  $X$

Garfield: „We never predicted that people would turn this into an evaluation tool for giving out grants and funding.“ From: Richard Monastersky (2005), *The Number That's Devouring Science* *The Chronicle of Higher Education*

## □ h-index: calculation

a scientist has index  $h$  if  $h$  of  $N$  papers have at least  $h$  citations each, and the other  $(N - h)$  papers have less than  $h$  citations each

# citation impact: a bunch of critiques

---

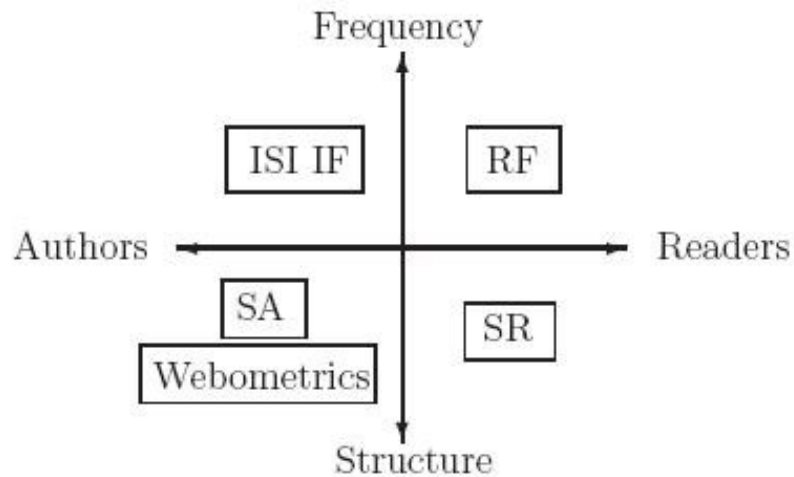
- ❑ restricted scope, exclusion of many publication types
- ❑ based exclusively on journal citation report/ web of science
- ❑ language bias: items in english language are overrepresented within the database, so they reach higher citation scores
- ❑ jif focuses on journals: few articles evoke most citations
- ❑ jif discriminates disciplines with lifecycles of scientific information > 2 years
- ❑ commixture of quality and popularity

# impact measures: a categorisation

---

- citation based measures
  - author centred
  - delayed measurement: at the first in the following generation of publications
  - mostly: impact of an separate object is not described
  
- usage based measures
  - reader centred
  - measuring: on-the-fly and consecutive
  - impact of a separate object can be described
  - automatised measurement possible

# impact measures: a categorisation, pt. II



**ISI IF = Journal Impact Factor**

**RF = Reading Factor**

**SA = Structure Author**

- based on networks built by authors and their activities, e.g. Google PageRank, citation graphs, webometrics

**SR = Structure Reader**

- based on document usage and its contextual information, e.g. Recommenders, download graphs

Bollen, J. et al. (2005): Toward alternative metrics of journal impact: A comparison of download and citation data. In: Information Processing and Management 41(6): S. 1419-1440.

Preprint Online: <http://arxiv.org/abs/cs.DL/0503007>



# usage based impact: standardisation?

---

- **COUNTER**, <http://www.projectcounter.org/>
- **LogEc**, <http://logec.repec.org/>
- **International Federation of Audit Bureaux of Circulations (IFABC)**, <http://www.ifabc.org/>

# Usage based impact: standardisation?

---

- the models mentioned differ in many respects
  - detection and elimination of non-human access (robots, automatic harvesting)
  - definition of double click intervals
  
- general problems
  - ignorance of context information
  - detection of duplicate users
  - detection of duplicate information items
  - ignorance of philosophical questions like: what degree of similarity makes two files the same document?

# alternative impact measures: conclusion

---

- ❑ alternative impact measures (in the form of usage based measures) can be mould
- ❑ but: very little standardisation
- ❑ promising, but complex examples/models like MESUR, <http://www.mesur.org/MESUR.html>
- ❑ requirement: sophisticated infrastructure to generate and exchange interoperable usage information within a network of several different servers

# Open Access Statistics

---

- funder: German Research Foundation (ger: Deutsche Forschungsgemeinschaft) DFG, <http://www.dfg.de>
  
- project partners:
  - Georg-August-University Göttingen (State- and University Library)
  - Humboldt-University Berlin (Computer- and Mediaservice)
  - Saarland University (Saarland University and State Library)
  - University Stuttgart (University Library)
  
- 07/2008 – 02/2010
  
- <http://www.dini.de/projekte/oa-statistik/english/>

# Open Access Statistics: motivation

---

- open access publications are often excluded from citation based impact measures
  - repository documents by definition
  - articles in open access journals due to their short citation history and often also due to their language
- citation based impact measures are revealing several deficiencies
- citation based impact measures should be complemented by usage based impact measures
  - because a multi-faceted approach could remedy some of their deficiencies
  - because the latter ones could create an incentive to use open access services
- it needs a project to establish the required infrastructure

# Open Access Statistics: aims

---

- ❑ implementation of a network to collect, process and exchange usage information between different services
- ❑ usage information should be processed according to the standards of COUNTER, LogEc and IFABC
- ❑ development of additional services for repositories
- ❑ development of implementation guidelines
- ❑ initially formulated by the Electronic Publishing working group of DINI (Deutsche Initiative für Netzwerkinformation / German Initiative for Network Information)

# Open Access Statistics: associated projects

---

- Open Access Statistics addresses usage description
- Open Access Citation address the issue of tracking citations between electronic publications
- Open Access Network
  - intends to build a network of repositories
  - will bundle the results of Open Access Citation and Open Access Statistics in one user interface
  - offers services for Open Access Citation and Open Access Statistics, e.g. deduplication of documents (based on a asymmetric similarity of fulltext documents)

# Open Access Statistics: background

---

- data pools at the partner institutions
  - open access repositories
  - linkresolver
  - licence controlling servers
  
- aggregation of usage information/ usage events from each single data pool in a central service provider
  - including deduplication
  - including processing according to the standards mentioned
  
- services provided by the central service provider
  
- usage data will be retransferred to distributed local repositories and to the Open Access Network service



# Open Access Statistics: example

---

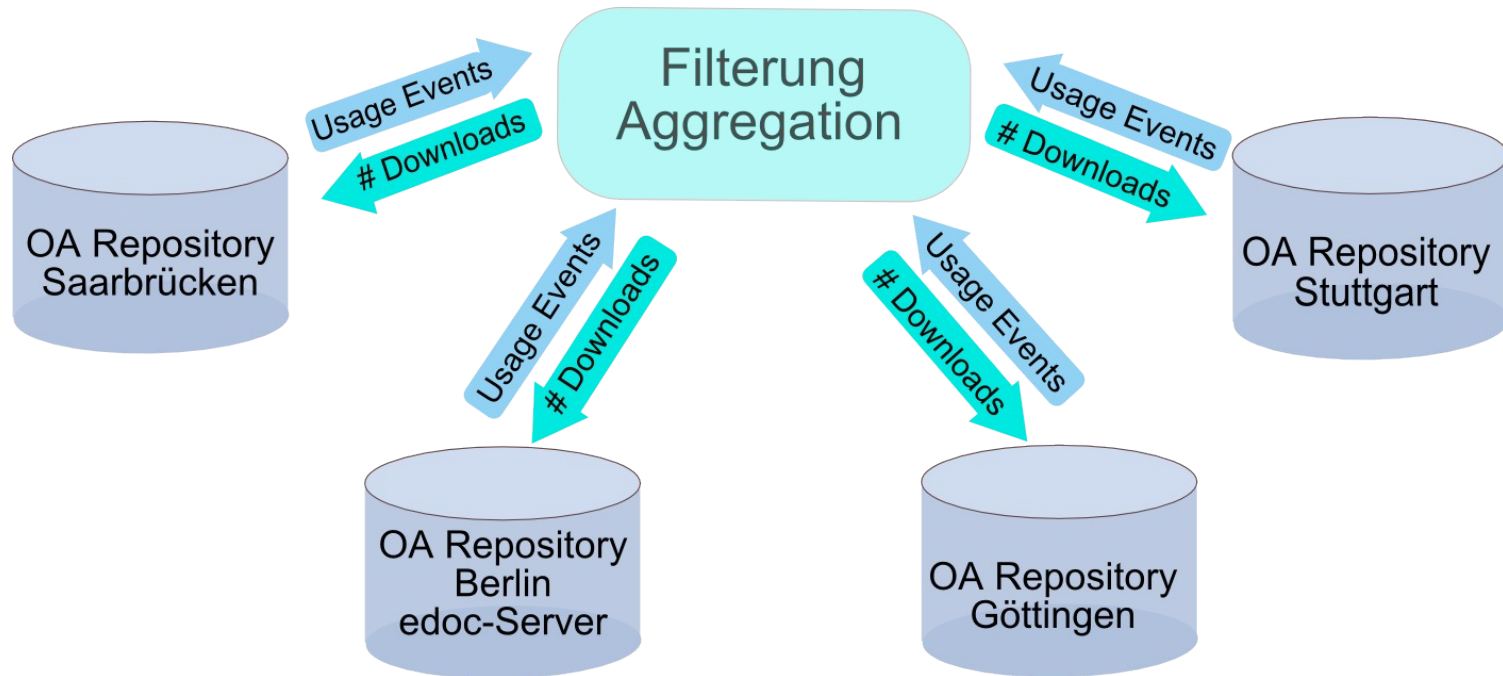
data provider (services x, y, z)

- generate logs about document usage
- pseudonymise user information (IP-addresses)
- process usage information (adds unique document ID, transforms data into OpenURL ContextObjects, ...)
- transmit the information via OAI-PMH to the service provider

service provider

- receives the information
- deduplicates documents and users
- computes usage statistics according to the standards mentioned
- delivers the information to external services (search engines, etc.) and to the data provider x, y, z that generated the logs

# Open Access Statistics: background



# Open Access Statistics: data provider

---

requirements

- a defined web server configuration
- local processing of the web server logs
  - pseudonymisation
  - isolation of the local document identification
  - ...
- packing of the OAI-PMH-container/ OpenURL-ContextObjects-container
  - referrent
  - reffering entity
  - requester
  - servicetype
  - resolver
  - referrer

# Open Access Statistics: data provider

---

retransfer of processed information to the local repository

- ▣ protocol: OAI-PMH
- ▣ syntax: XML
- ▣ resolution: to be discussed (month?, weeks?, days?)
- ▣ granularity: fulltexts

# Open Access Statistics: some lessons learned

---

linkresolvers are rarely offering suitable information

- external services (ovid) don't offer usage information
- SFX-logs are very heterogenous
  - target may be a splash page or a fulltext
- hardly any information about open access documents

document deduplication seems difficult

- a given document may have more than one IDs  
cause: multiple fulltext deposit on several repositories
- a given document may have several splash pages on different servers pointing at one fulltext on one single server  
cause: metadata harvesting
- ...

# Open Access Statistics: usage scenarios

---

data may be used

- ❑ from an user perspective as a criterion to estimate the relevance of a document (e.g. rankings)
- ❑ from an author perspective as an indicator for the dissemination of a concept
- ❑ from a service provider perspective:
  - as additional metadata for search engines, databases ...
  - as a recommender service
- ❑ from a repository perspective:
  - as a recommender service
  - as additional metadata for users

# Open Access Statistics: repository integration



English

Lizenz

ResearchPa  
URN: urn:n  
URL: http://

How to find

Koschorrec

Originalve  
pdf-Format

Dokument

Dokument  
Scholar su

SWD-Schle

Freie Schli  
(Englisch):

Institut:

DDC-Sache

Dokumenta

Sprache:

Erstellungs

Publikations

Kurzfassung auf

Englisch:



In screening of libraries derived by expression cloning, expression of active proteins in *E. coli* can be limited by formation of inclusion bodies. In these cases it would be desirable to enrich gene libraries for coding sequences with soluble gene products in *E. coli* and thus to improve the efficiency of screening. Previously Wilkinson and Harrison showed that solubility can be predicted from amino acid composition (Biotechnology 1991, 9(5):443-448). We have applied this analysis to members of the alpha/beta hydrolase fold family to predict their solubility in *E. coli*. alpha/beta hydrolases are a highly diverse family with more than 1800 proteins which have been grouped into homologous families and superfamilies.

Results:

The predicted solubility in *E. coli* depends on hydrolase size, phylogenetic origin of the host organism, the homologous

# Open Access Statistics: repository integration

Georg-August-Universität Göttingen Impressum | Kontakt | Feedback

**GoeScholar** > Naturwissenschaften, Mathematik und Informatik > Fakultät für Chemie > Zeitschriftenartikel - Fakultät Chemie >

Suche     
Erweiterte Suche

**Browsen**  
» Fakultäten  
» Zentren  
» Titel  
» Autor  
» Jahr

**Mein GoeScholar**  
» Dokument(e) anmelden  
» Checkliste  
» Meine Publikationsliste  
» Benutzerprofil  
» **DEMO**

**Infos & Hilfe**  
» Hilfe & FAQ  
» Rechtseinräumung

**Login:**  
GWDG-User:   
Passwort:

**Verlinken Sie auf bzw. zitieren Sie dieses Dokument mit der folgenden permanenten URL:**  
<http://resolver.sub.uni-goettingen.de/purl?goescholar/3377>

**Nutzungsstatistik**

COUNTER IFABC LogEc

| Datum      | Aufrufe |
|------------|---------|
| 28.09.2009 | 20      |
| 30.09.2009 | 13      |
| 01.10.2009 | 10      |
| 04.10.2009 | 8       |

Summe letzte 7 Tage: 89

Zeitraum: **7 Tage** 30 Tage 90 Tage

**Serie/report Nr.:** Organometallics, Vol. 12, 4, 1193-1200

**Zusammenfassung:** The set of alkali metal solid-state structures of Ph<sub>3</sub>CM-nL [M = K, Rb, Cs; L = (ligands) PMDTA (N,ZV,"JV"JV"-pentamethyldiethylenetriamine), THF (tetrahydrofuran)] provides instructive comparisons. Ph<sub>3</sub>CK-THF-PMDTA (1) crystallizes as a monomeric contact ion pair: the K<sup>+</sup> cation is symmetrically  $\eta$ -coordinated to one of the phenyl rings, but not to the deprotonated central carbon. Both [Ph&Rb-PMDTA], (2) and P~CCWPMDTA(3.) form one-dimensional polymers and eschew THF. The Rb cations in 2 bridge the triphenylmethyl moieties by +x"%"xmtion to separate phenyl rings. This gives rise to a zigzag chain. In 3, each Cs cation also bridges two carbanions, but in a somewhat different fashion. While Cs<sup>+</sup> is located rather symmetrically (119 above the phenyl ring of one trityl moiety, a,"propeller"-like coordination to a second trityl anion



# Open Access Statistics: additional information

---

- ❑ open access statistic will offer modules for OPUS- and DSpace-based repositories, other products can be configured easily
  - ❑ Nutzungsstatistiken elektronischer Publikationen. DINI-Schriftenreihe. DFG-Projekt Open Access-Statistik (OA-S) und DINI-Arbeitsgruppe „Elektronisches Publizieren“. Online: <http://nbn-resolving.de/urn:nbn:de:kobv:11-100101174> (to be translated)
- ❑ Open Access Statistics workshop: 21.01.2010
  - ❑ [http://www.dini.de/veranstaltungen/workshops/oa-statistik-was\\_zaeHLT/](http://www.dini.de/veranstaltungen/workshops/oa-statistik-was_zaeHLT/)
- ❑ online questionnaire on features in digital repositories
  - ❑ <http://oas.sulb.uni-saarland.de/fragebogen-english.php>
- ❑ online demo
  - ❑ <http://oa-statistik.sub.uni-goettingen.de/statsdemo>
- ❑ website with further information about the workshop, technical specifications
  - ❑ <http://www.dini.de/projekte/oa-statistik/english/>

# Open Access Statistics: further plans

---

## Open Access Statistics II?

possible focus:

- internationalisation
- opening the network to other contributing repositories
- opening the network to other services (e.g. journals)
- evaluation of metrics more complex than the calculation of pure usage frequencies
- ...

# Open Access Statistics: cooperation

---

- ❑ SURFSure  
Statistics on the Usage of Repositories
- ❑ COUNTER  
Counting Online Usage of Networked Electronic Resources
- ❑ PIRUS  
Publisher and Institutional Repository Usage Statistics
- ❑ NEEQ  
Network of European Economists Online
- ❑ PEER  
Publishing and the Ecology of European Research
- ❑ OAPEN  
Open Access Publishing in European Networks


**Thanks for your attention!**

**And thanks to my colleagues:**

**Bettina Bauer**

**Daniel Metje**

**Björn Mittelsdorf**



Université Lille 3: International Symposium on  
„Academic Online Ressources : Assessement and Usage“  
26.11.2009

Initiated by:



**Ulrich Herb**

Saarland University and State Library, Germany  
u.herb@sulb.uni-Saarland.de

Funded by:

Deutsche  
Forschungsgemeinschaft

**DFG**